Filozofia w Praktyce
Tom 5 (2019), Artykuł 10

Kogo potrąci samochód autonomiczny? Moralność pojazdów bezzałogowych

– Alicja Krasnowska –

Biorąc pod uwagę postęp technologiczny, samochody autonomiczne, nazywane również bezzałogowymi, już wkrótce mogą zostać wprowadzone do powszechnego użycia na całym świecie. Mimo że jedną z najważniejszych zalet rozpowszechnienia się takich samochodów mogłoby być zmniejszenie liczby wypadków drogowych (na przykład tych spowodowanych zmęczeniem lub nieuwagą kierowcy, czy też niezastosowaniem się do zasad ruchu drogowego), to w niektórych sytuacjach kolizje z ich udziałem wydają się trudne do uniknięcia. Samochody autonomiczne muszą być zatem wyposażone w odpowiedni algorytm działania w podobnych okolicznościach. Powstaje pytanie, jak ten algorytm zaprojektować, a w szczególności, jakie zasady etyczne należy wziąć pod uwagę. Decyzje, które kierowcy w razie wypadku podejmują instynktownie w ułamkach sekund, tutaj będą obliczane przez program, który z założenia może wykonać w bardzo krótkim czasie szereg precyzyjnych operacji w celu rozstrzygnięcia, jaką trasę powinien obrać samochód. Oczekujemy, że wynik tak dokładnych obliczeń zawsze będzie wskazywał na optymalne rozwiązanie. Kwestia rozstrzygnięcia, czym konkretnie to optymalne rozwiązanie powinno się charakteryzować, pozostaje jednak nieoczywista, zwłaszcza w sytuacjach wyboru mniejszego zła, kiedy nie zredukujemy szkód do zera. Rozwiązań można szukać między innymi w intuicjach ludzi, w neuronaukach, a także w teoriach użyteczności zaczerpniętych z nauk ekonomicznych.

Projekt „Moral Machine"

Jednym z ciekawszych pomysłów na rozwiązanie omawianego problemu było sprawdzenie intuicji etycznych osób z całego świata, aby na ich podstawie wysnuć wnioski o najważniejszych oczekiwaniach co do moralności samochodów autonomicznych. Badanie takie zostało zrealizowane w projekcie „Moral Machine" prowadzonym na MIT (Awad et. al. 2018). Autorzy projektu uznali, że nie tylko etycy powinni brać udział w ustalaniu zasad postępowania dla aut bezzałogowych i należy uwzględnić zdanie szerszej grupy potencjalnych użytkowników takich samochodów. Eksperyment przeprowadzono on-line na ogromnej próbie: uczestniczyło w nim prawie 40 milionów respondentów z całego świata. Każdy badany podejmował decyzje w trzynastu losowych scenariuszach, w których należało wybrać trasę samochodu autonomicznego w sytuacji nieuniknionego wypadku. Taki scenariusz mógł zawierać przykładowo możliwość poświęcenia pasażerów albo grupy przechodzących przez ulicę na zielonym świetle pieszych. Jednym z najważniejszych aspektów, na jakie zwrócono uwagę przy analizie wyników, były różnice międzykulturowe. Respondenci z krajów południowych preferowali ratowanie

kobiet, uczestnicy z państw kultury wschodniej – ratowanie pieszych, badani z państw zachodnich wybierali częściej niepodjęcie żadnego działania, to znaczy niezmieniania trasy samochodu. Co ciekawe, priorytet przyznany pieszym przez uczestników z Polski był najniższy w Europie i osiągnął poziom poniżej światowej średniej.

Mimo takich różnic badaczom udało się wyodrębnić trzy najważniejsze zasady, które mogłyby potencjalnie stanowić podstawę szeroko akceptowanego przez społeczeństwo kodeksu postępowania dla aut bezzałogowych. Po pierwsze, należy, kiedy to tylko możliwe, oszczędzać życie ludzkie. Po drugie, trzeba dążyć do ocalenia jak największej liczby osób (taka zasada byłaby zgodna z utylitarnym podejściem do podobnych dylematów). Po trzecie, preferowane jest ratowanie osób młodszych, w szczególności małych dzieci. Ostatnią zasadę, która wyraża co prawda powszechną i silną intuicję występującą u osób badanych, warto porównać z raportem „Automated and Connected Driving" niemieckiej Komisji Etycznej z czerwca 2017 roku. Jest to jedyny sformułowany do tej pory oficjalny dokument zawierający wytyczne dla osób odpowiedzialnych za etykę zautomatyzowanego ruchu drogowego. Zawarty jest w nim punkt dotyczący zakazu kierowania się wiekiem, płcią i innymi cechami osobowymi przy podejmowaniu decyzji, kogo narazić na szkodę w wypadku. Na potencjalne niebezpieczeństwa związane z kierowaniem się wymienionymi cechami zwraca uwagę Hin-Yan Liu (Liu: 2016), prezentując dystopijny scenariusz, który mógłby spełnić się, gdyby zaprogramować samochody autonomiczne do masowego dyskryminowania konkretnych grup uznanych z różnych powodów za mniej przydatne dla społeczeństwa. Liu słusznie zauważa, że o ile w pojedynczych przypadkach da się usprawiedliwić wybory oparte na cechach osobowych, to kiedy spojrzymy na problem całościowo, jak na całą sieć osób uczestniczących w wypadkach, okaże się, że mamy do czynienia ze zjawiskiem niedopuszczalnym: pewne grupy są obciążane ogromnym ryzykiem, a inne znacznie faworyzowane.

Neuroobrazowanie: logika i emocje

Kolejne podejście, które moglibyśmy zastosować poszukując odpowiedniego kodeksu moralnego dla samochodów autonomicznych, znów dotyczy badania ludzi, jednak tym razem pomysł opiera się na wykorzystaniu badań z neuroobrazowaniem. Zaobserwowanie u badanych, jakie obszary mózgu uczestniczą w rozwiązywaniu dylematów etycznych, mogłoby pomóc nam w zrozumieniu tego mechanizmu, a nawet zaimplementowaniu podobnego w samochodach bezzałogowych. Przykładem eksperymentu, w którym podczas podejmowania trudnych decyzji moralnych obserwowane były pobudzenia różnych obszarów kory mózgowej, jest badanie przeprowadzone z wykorzystaniem fMRI i dwóch scenariuszy wypadków, w których rozstrzygali badani (Greene 2001). Okazało się, że przy różnych wersjach zaprezentowanych sytuacji, aktywne były różne części kory mózgowej, odpowiedzialne odpowiednio za procesy poznawcze oraz za reakcje emocjonalne. Autor eksperymentu wyprowadził wniosek, że w rozwiązywanie problemów etycznych u ludzi ingerują emocje (niekoniecznie musi być to coś negatywnego), a wyraźne różnice w odpowiedziach badanych wynikały właśnie z faktu wystąpienia lub braku reakcji emocjonalnej. Spostrzeżeniem istotnym dla kwestii programowania samochodów jest to, że zaprogramowanie ich na wzór ludzi nie będzie łatwe, między

innymi ze względu właśnie na wchodzące w grę emocje. Oczywiście można by przyjąć, że decyzje podejmowane bez udziału emocji uznajemy jednak za w jakimś sensie lepsze, bardziej logiczne czy racjonalne, a zatem zaprojektujemy algorytm na wzór postępowania badanych w tych przypadkach, w których nie wystąpiła reakcja emocjonalna. To z kolei wskazywałoby na podejście zbieżne z poglądem utylitarnym, a więc ratowanie większej liczby osób, nawet gdy oznacza to w zamian poświęcenie jednostki. Z drugiej strony warto zastanowić się, czy łatwe odrzucenie roli emocji w rozwiązywaniu dylematów nie oznaczałoby, że zaprojektowany w ten sposób system byłby trudny do zaakceptowania dla niektórych użytkowników. Pomysł odrzucenia osobowego podejścia do człowieka na rzecz racjonalnej kalkulacji, traktującej osoby uczestniczące w wypadku jako nierozróżnialne przedmioty, mógłby wywołać sprzeciw właśnie ze względu na to, że we wspomnianym badaniu emocje okazały się u ludzi ważnym czynnikiem przy podejmowaniu decyzji w dylematach moralnych.

Konsekwencjalizm i użyteczność

Kolejne ciekawe podejście przedstawia Vanessa Schäffner (Schäffner: 2018), która opiera swoją koncepcję na zasadach pochodzących z nauk ekonomicznych. Autorka zwraca uwagę, że podejście konsekwencjalistyczne do rozstrzygania dylematów związanych z nieuniknionymi wypadkami można opisać w sposób matematyczny, ponieważ polega ono w uproszczeniu na maksymalizowaniu użyteczności przy pomocy analizy możliwych skutków danego wyboru. Decyzja zależy zatem od tego, która z możliwości generuje najwięcej korzyści najmniejszym kosztem. W funkcji kosztu możemy uwzględnić istotne różnice między stratami ludzkimi a materialnymi oraz kary związane z naruszeniem zasad ruchu drogowego. Takie podejście nie jest jednak pozbawione wad, związanych choćby z trudnościami w wycenie, jak wielką stratą jest śmierć człowieka, przypisywaniu właściwych wag potencjalnym stratom i zyskom, oraz uwzględnieniu w obliczeniach kontekstu sytuacji czy skutków długofalowych. Rozwiązaniem, jakie proponuje autorka, jest koncepcja oparta na utylitaryzmie negatywnym (rozumianym tu najogólniej jako pogląd mówiący, że z moralnego punktu widzenia lepiej dążyć do minimalizacji cierpienia niż maksymalizacji zadowolenia). Pomysł przedstawiony przez Schäffner polega zatem na unikaniu ofiar śmiertelnych, których wystąpienie autorka uznała w swoich rozważaniach za największe zło w sytuacji kolizji, to znaczy najgorszy możliwy skutek wypadku, oraz ochronie tych uczestników ruchu, którzy są najsłabiej chronieni przed kolizją (na przykład pieszych). Dzięki takiemu podejściu minimalizujemy straty, które ponoszą ludzie. Ucierpią tylko ci uczestnicy ruchu, którzy mają największą szansę uniknięcia najpoważniejszych skutków wypadku. Przyznanie pierwszeństwa tym, którzy będą w największym niebezpieczeństwie, wydaje się szlachetnym podejściem do problemu, jednak tylko pod warunkiem, że zadbamy również o bezpieczeństwo tych, którzy z punktu widzenia szans na przeżycie są w sytuacji bardziej komfortowej, czyli pasażerów. Projektanci samochodów autonomicznych mogą brać pod uwagę wbudowanie dowolnie wielu środków ostrożności takich jak poduszki, czy nawet awaryjne katapulty w siedzeniach, które zastosowane razem pozwolą skupić się w algorytmie na ochronie pozostałych zagrożonych jednostek.

Chociaż rozważając problemy związane z zaprogramowaniem samochodów autonomicznych rozpatruje się bardzo drastyczne scenariusze, to nie robi się tego przecież z powodu przewidywania, że takie wypadki będą zdarzać się często, lecz aby zapewnić maksymalne możliwe bezpieczeństwo w razie takiego zdarzenia. Mimo moralnych wyzwań związanych z wprowadzeniem aut bezzałogowych są one przyszłościowym rozwiązaniem, które ma szansę pozwolić na znaczne zredukowanie liczby wypadków drogowych, zwłaszcza jeśli wykorzystując ogromną ilość danych, jakimi będą dysponowały, udałoby się wyposażyć je w system rozpoznawania sytuacji niebezpiecznych, zanim w ogóle do nich dojdzie. Mogłoby to sprawić, że rozważane przez badaczy skrajne dylematy stałyby się w praktyce naprawdę rzadkością.

Literatura

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. (2018). *The Moral Machine experiment*. „Nature" 563: 59–64.

Liu H.-Y. (2016). *Structural Discrimination and Autonomous Vehicles: Immunity Devices, Trump Cards and Crash Optimisation*. W: *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016 / TRANSOR 2016*. J. Seibt, M. Nørskov, S. Schack Andersen (red.). IOS Press.

Schäffner V. (2018). *Caught Up in Ethical Dilemmas: An Adapted Consequentialist Perspective on Self-Driving Vehicles*. W: *Envisioning Robots in Society – Power, Politics, and Public Space: Proceedings of Robophilosophy 2018 / TRANSOR 2018*. M. Coeckelbergh, J. Loh, M. Funk (red.). IOS Press.

Alicja Krasnowska – studentka III roku kognitywistyki na Uniwersytecie Warszawskim

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498).